Lokalisierung von Schrift in komplexer Umgebung

KARL-HEINZ STEINKE[1]

Das Forschungsprojekt „Herbar Digital" startete 2007 mit dem Ziel der Digitalisierung des Bestands von mehr als 3,5 Millionen getrockneter Pflanzen bzw. Pflanzenteile auf Papierbögen (Herbarbelege) des Botanischen Museums Berlin. Die Aufgabe des Autors ist die Analyse der hochaufgelösten Bilder mit 10400 Zeilen und 7500 Spalten. Die Herbarbelege können außerdem unterschiedliche Objekte enthalten wie Umschläge mit zusätzlichen Pflanzenteilen, gedruckte oder handgeschriebene Etiketten, Farbtabellen, Maßstäbe, Stempel, Barcodes, farbige „Typus-Etiketten" und handschriftliche Anmerkungen direkt auf dem Beleg. Die schriftlichen Anmerkungen, insbesondere in Handschrift, sind von besonderem Interesse. Kommerzielle OCR-Software kann oftmals Schrift in komplexen Umgebungen nicht lokalisieren, wie sie häufig auf den Herbarbelegen vorliegt, auf denen Schrift zwischen Blättern, Wurzeln und anderen Objekten angeordnet ist. Im folgenden wird eine Methode vorgestellt, die es ermöglicht, Schriftpassagen im Bild automatisch zu finden.

1 Einleitung

Kommerzielle OCR-Programme wie z.B. Finereader 9.0 oder Omnipage 16 sind in der Lage, hochqualitativen gedruckten Text weitestgehend korrekt zu erkennen. Allerdings gibt es noch erhebliche Defizite, Texte in komplexen Dokumenten mit eingebetteten Bilddaten oder mit anderen Objekten korrekt zu verarbeiten. Die vorliegenden Herbarien stellen eine komplexe Umgebung für die verschiedenen Texte (Druck- und Handschriften) dar, außerdem sind zusätzliche Objekte vorhanden wie Stempel, Barcode, Maßstäbe, Farbtafeln, Tüten usw., die es sehr schwierig machen, Textstellen im Bild zu lokalisieren. In der Literatur gibt es einige Ansätze, Texte in Bildern zu lokalisieren, wie z.B. auf Buchumschlägen, Scheckformularen, in farbigen Anzeigen, Videobildern, Internetbildern oder allgemeinen Farbbildern, die autonome Roboter mit ihrer Videokamera aufnehmen. In den meisten Fällen handelt es sich dabei um Druckschrift, die erkannt werden soll. Bei den vorliegenden Herbarien stellt jedoch das Gemisch von Handschriften unterschiedlicher Schreiber, Druckschrift in unterschiedlichen Größen und Formen, gestempeltem Text und Barcodeinformationen eine besondere Herausforderung dar.

2 Beschreibung der Methode

Das Originalfarbbild (siehe Abb. 2) mit 600 dpi wird auf ein Grauwertbild mit 150 dpi reduziert, um die Rechenzeit in einem vertretbaren Rahmen zu halten. Es wird ausgegangen von horizontal geschriebenen Texten mit nur einer geringen Abweichung von maximal 10 Grad von der Horizontalen. Um Schriftpassagen im Bild automatisch zu finden, muss man sich die Eigenschaften von Schrift zunutze machen. Handschrift besteht im Wesentlichen aus

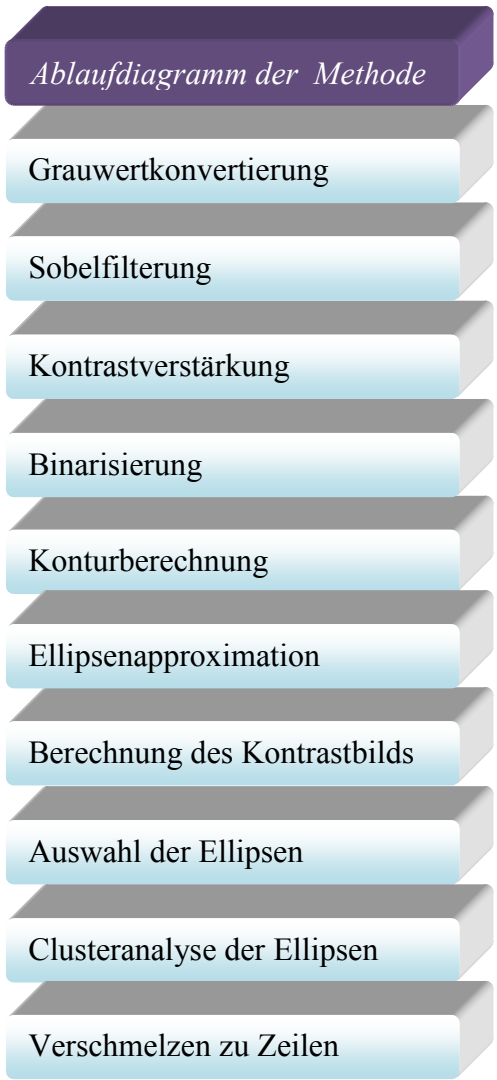
1) Karl-Heinz Steinke, Fachhochschule Hannover, Ricklinger Stadtweg 120, 30459 Hannover;
E-Mail: karl-heinz.steinke@fh-hannover.de

Tab. 1: Sobeloperator

-1	0	1
-2	0	2
-1	0	1

Auf- und Abwärtsbewegungen, die sich von links nach rechts in einer Schreibzeile fortbewegen. Die entstehenden vertikalen Linien lassen sich gut mit dem Sobeloperator (siehe Tab. 1) durch Faltung des Bildes mit einem 3*3 Fenster gewinnen. Der Sobeloperator kombiniert Gausssche Weichzeichnung und partielle Ableitung nach x, so dass das Ergebnis einigermaßen robust gegenüber Rauschen ist (siehe Abb. 3). Durch die Sobel-Filterung wird zusammenhängende Handschrift und Druckschrift zerlegt in kleine nahezu vertikale Schriftsegmente, die mit ihrer Schräglage gut die Schriftneigung wiedergeben.

Tab. 2: Ablaufdiagramm der Methode



Um die Segmente deutlich vom Hintergrund abzuheben, wird eine Kontrastverstärkung mit anschließender Binarisierung durchgeführt. Man erhält vertikal ausgerichtete Linienelemente, deren Konturen berechnet werden (siehe Abb. 4). Um die Merkmale der Objekte nämlich

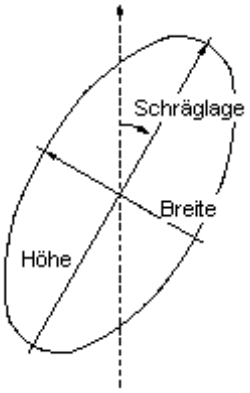


Abb. 1: Ellipsenparameter

Höhe, Breite, Winkel und Schwerpunkt (x,y), auf einfache Weise zu erhalten, bietet sich eine Ellipsenapproximation nach der Methode der kleinsten Fehlerquadrate an. Die Merkmale der Ellipsen, d.h. Höhe, Breite, Winkel und Schwerpunkt werden für alle Ellipsen berechnet und entsprechen etwa den Merkmalen der Objekte. Wie aus Abb. 5 zu erkennen ist, stellt die Gesamtheit aller Ellipsen unterschiedliche Objekte wie Pflanzenteile, Stempel, Teile der Farbtabellen aber auch Schriftanteile dar. Es sind nun diejenigen Ellipsen herauszufiltern, die Schrift darstellen. Dazu müssen mit heuristischen Methoden aus allen Objekten nur diejenigen ausgewählt werden, die „Schrifteigenschaften“ besitzen (siehe Tab. 3).

Abb. 2: Originalbild

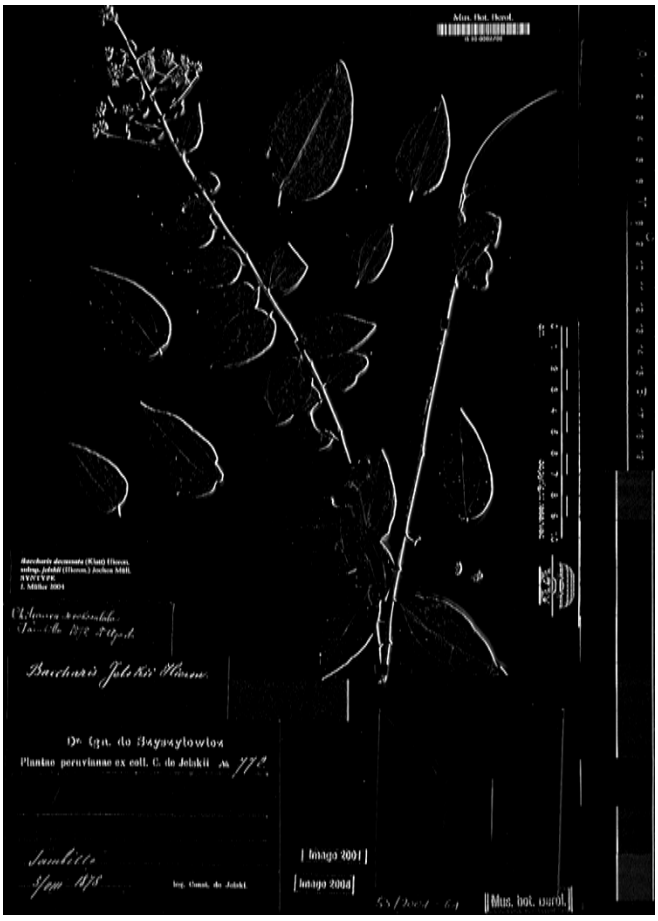


Abb. 3: Sobelfilterung

Abb. 4: Konturen

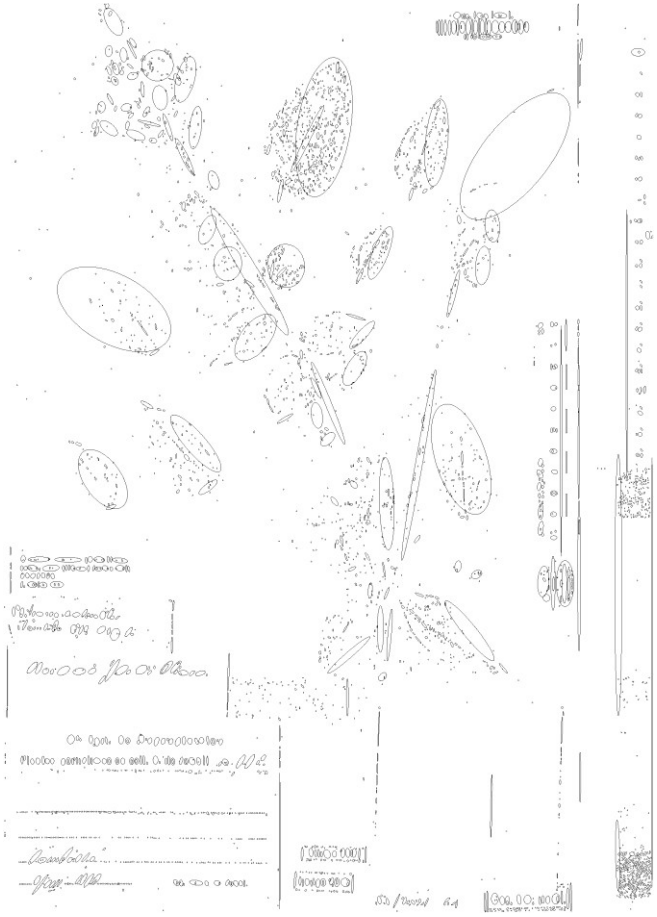


Abb. 5: Ellipsenapproximation

Schrift besitzt einen hohen Kontrast und eine horizontale Zeilenstruktur. Wegen der schnellen Abfolge von schwarzen Strichen auf weißem Hintergrund ist insbesondere der horizontale Kontrast besonders ausgeprägt. Unter 2.1 wird ein Verfahren erläutert, um diesen Kontrast für jede Position im Bild zu bestimmen. Zwischen den Schreibzeilen fällt der hohe horizontale Kontrast schnell ab, da der Zeilenzwischenraum fast nur aus weißem Hintergrund besteht. Natürlich sollen die Schriftobjekte eine definierte Höhe besitzen (Schrifthöhe), um als schriftwahrscheinlich zu gelten. Dazu wird ein Bereich definiert, in dem sich die Schrifthöhe bewegen darf. Außerhalb des Bereichs liegende Objekte werden abgelehnt. In einer nahen Umgebung der Objekte soll es „schriftartig“ aussehen, d.h. die Textur hat bestimmte statistische Merkmale. Unter 2.2 wird der statistische Ansatz erläutert, der auch zur Unterscheidung verschiedener Schriftarten beiträgt. Wegen der Zeilenstruktur haben Schriftelemente normalerweise Nachbarn in der gleichen Zeile. Benachbarte Objekte werden zu Schriftzeilen verschmolzen, was durch einen Clusteranalyseansatz unter 2.3 gelingt.

Tab. 3: Eigenschaften von Schrift

Eigenschaften von Schrift
Hoher Kontrast
Zeilenstruktur
Größe im Toleranzbereich
Umgebung ist schriftartig
Nachbarn in gleicher Zeile

2.1 Berechnung des Kontrastbilds

Es wird für jeden Bildpunkt in einer quadratischen Umgebung (z.B. 30*30) der Kontrast berechnet und als Helligkeitswert im Kontrastbild (siehe Abb. 7) dargestellt. Es fällt auf, dass nicht nur die Schriftbereiche im Bild einen hohen Kontrast aufweisen, sondern auch Objekte wie Blattränder, Farbtabellen und Zentimetermaße. Da insbesondere der horizontale Kontrast die Schrift von anderen Objekten unterscheidet, wurde versucht, nur die Umgebung in der Zeile auszuwerten (z.B. 30*1). Neben einer Einsparung von Rechenzeit zeigt Abb. 8 auch ein verbessertes Ergebnis. In Abb. 9 wird ein noch besseres Ergebnis über ein schnellen Algorithmus erreicht, der die Anzahl der Farbwechsel von hell nach dunkel in einer Zeilenumgebung berechnet. Man erkennt, dass z.B. die Pflanze und die Farbtabelle weitgehend unterdrückt werden und hauptsächlich Schriftbereiche und der Barcode stark in den Vordergrund treten. Die Schriftzeilen machen sich nun als helle Streifen im Bild bemerkbar. So haben helle Bildbereiche in Abb. 9 eine besonders hohe Wahrscheinlichkeit Schrift zu enthalten. Aus allen Ellipsen werden nur diejenigen ausgewählt, die in hellen Bereichen des Farbwechsel-Kontrastbildes liegen und zusätzlich nach oben und unten ein schneller Abfall der Helligkeit (Zeilenzwischenraum) erfolgt.

Abb. 6: Original

Abb. 7: Kontrast

Abb. 8: Zeilenkontrast

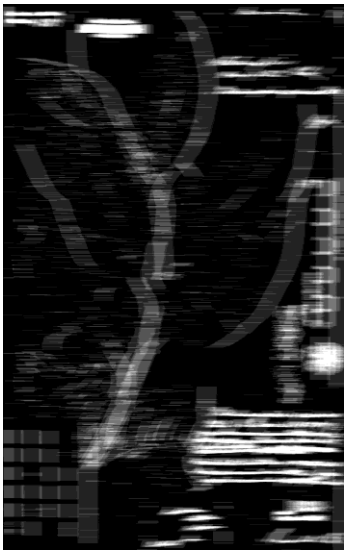


Abb. 9: Farbwechsel

2.2 Statistischer Ansatz

Um die Umgebung eines als Schrift in Frage kommenden Objekts zu analysieren, wurde ein statistischer Ansatz gewählt. Zur Texturbeschreibung wird eine Lauflängenverteilung in acht verschiedenen Richtungen (siehe Abb. 10) berechnet, die sich durch das Bildraster ergeben. Die Lauflängen werden in jeder Richtung in 8 Merkmalvektoren einsortiert, so dass sich insgesamt ein Merkmalvektor mit 64 Komponenten ergibt. Mit diesen Merkmalen ist man in der Lage, nicht nur zu erkennen, ob es sich um Schrift handelt oder nicht. Man kann sogar feststellen, ob es sich um Druck- oder Handschrift handelt.

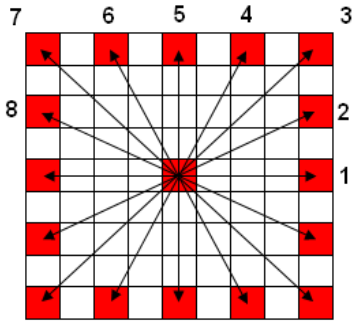


Abb. 10: Acht Richtungen

Abb. 11: Druckschrift

Abb. 12: Handschrift

In einer Lernphase können beispielsweise große und kleine Druckschriften oder verschiedene Handschriften sowie Stempel und Barcodes angelernt werden. In Abb. 17 sind die verschiedenen erkannten Muster unterschiedlich eingefärbt.

2.3 Clusteranalyse

Der Clusteranalyseansatz wurde in der Hoffnung gewählt, dass damit Textbereiche gut von Pflanzen und anderen Objekten separiert werden können. Es bietet sich an, die Ellipsen mit ihren 5 Parametern (Höhe, Breite, Winkel, Schwerpunkt(x,y)) als Punkte in einem 5-dimensionalen Vektorraum aufzufassen und zu Punktwolken (Clustern) zusammenzufassen. Die 5 Parameter können mit unterschiedlicher Gewichtung versehen werden. Um zusammenhängende Schrift in ein Cluster zu bekommen, liegt es nahe, die Zeilenähnlichkeit stärker zu gewichten als die Spaltenähnlichkeit. Beim eingesetzten k-means-Algorithmus wird die Anzahl k von Clustern vor dem Start festgelegt. Wählt man die Clusterzahl gleich der Anzahl der Textbereiche, arbeitet die Clusteranalyse recht gut (siehe Abb. 13). Es ist jedoch unbekannt, wie viele Textstellen im Bild vorhanden sind. Wie in Abb. 14 zu erkennen ist, verläuft die Clusteranalyse nicht zufriedenstellend, wenn die Clusteranzahl nicht mit der Anzahl der Textbereiche übereinstimmt. Beim Experimentieren mit den Gewichtsfaktoren lagen zusammengehörige Textbereiche oft in unterschiedlichen Clustern oder verschiedene Textbereiche wurden in ein Cluster

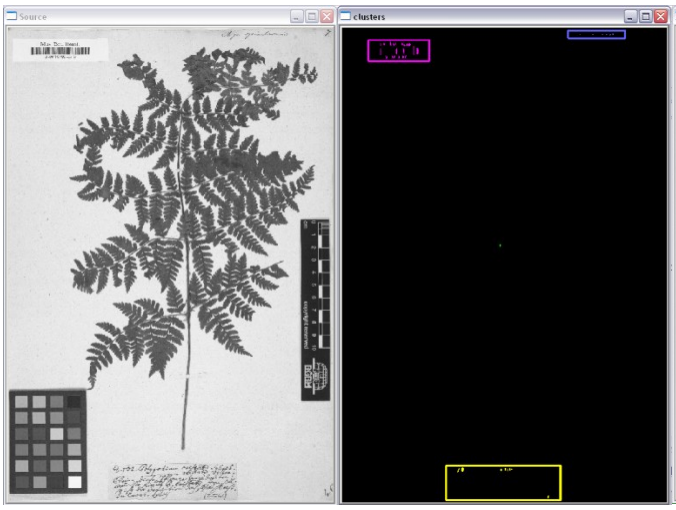


Abb. 13: Clusteranzahl entspricht Textbereichsanzahl

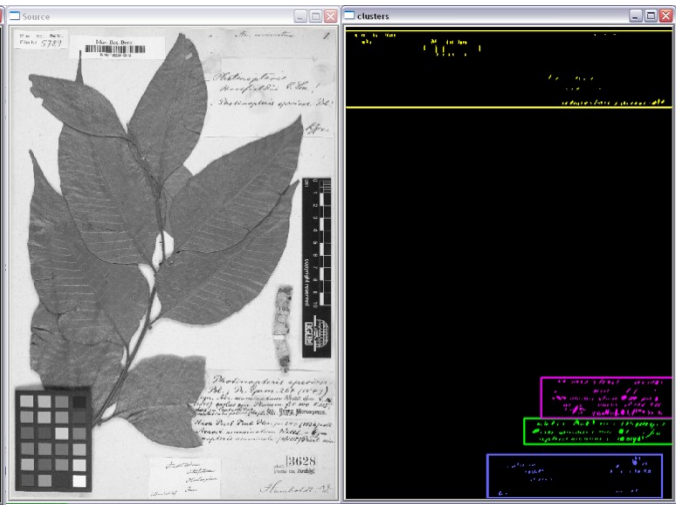


Abb. 14: Clusteranzahl kleiner Textbereichsanzahl

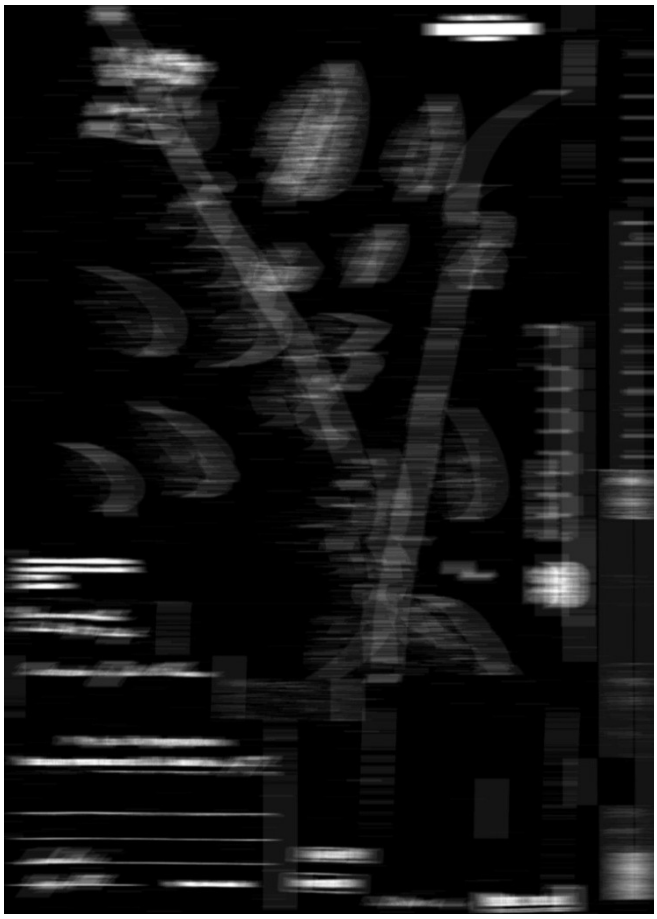


Abb. 15: Kontrastbild über Farbwechsel

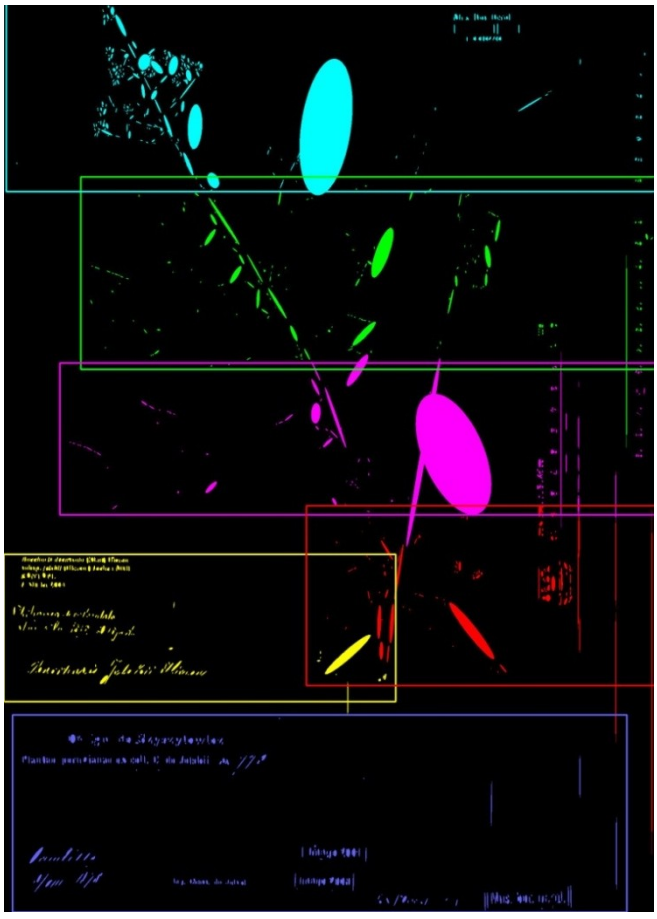


Abb. 16: Clusteranalyse

zusammengefasst. Aus diesem Grund wurde ein eigener, die Zeilenstruktur berücksichtigender Clusteralgorithmus entwickelt. Um die schriftartigen Objekte, die alle gewünschten „Schriftmerkmale“ besitzen, zu Textzeilen zusammenzufassen, versieht man jede Ellipse mit einem Label. Die Labels werden verschmolzen, wenn ihre Ellipsen in einer Zeile dicht nebeneinander liegen. Dieser Vorgang wird solange wiederholt, bis keine Verschmelzungen mehr möglich sind. In diesen Verschmelzungsprozess werden auch zunächst unsichere Kandidaten, die die Eigenschaft „hoher Kontrast“ nicht erfüllten, mit einbezogen. Auf diese Weise werden am Ende alle schriftartigen Objekte zu Zeilen verschmolzen, die dann aus dem Bild ausgeschnitten und weiter analysiert werden können. Man erkennt in Abb. 17, dass die Pflanze fast komplett herausgefiltert wird und auch die Farbtabelle und das Zentimetermaß völlig verschwindet. Übrig bleiben im Wesentlichen nur Druck- und Handschriften, sowie der Barcode und Stempel.

3 Ergebnisse und Ausblick

Als Datenmaterial sind 465 Bilder vorhanden mit unterschiedlichsten Pflanzen, gedruckten und maschinegeschriebenen Texten, handschriftlichen Eintragungen von unterschiedlichen

Abb. 17: Schriftzeilen

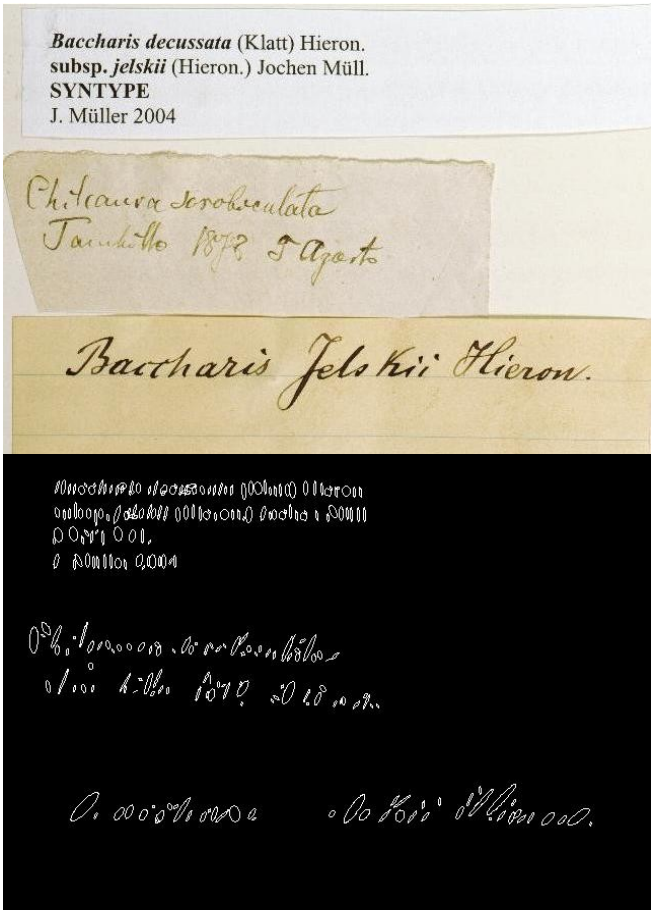


Abb. 18: Schrift mit Ellipsenanpassung

Schreibern u.a. auch von Alexander von Humboldt, Stempeln usw. Auch in Fällen, in denen hochwertige kommerzielle OCR-Programme Textzeilen nicht lokalisieren konnten, gab das obige Verfahren deutliche Hinweise auf Schriftzeilen. In einigen Fällen wurden allerdings auch Pflanzenbestandteile, insbesondere horizontal liegende Gräser, als mögliche Textregionen angegeben. Solche Fehler treten jedoch auch bei kommerziellen OCR-Programmen auf. So erkannte Omnipage 16 ein Pflanzenblatt in einem Bild als geschriebenen Text. Insgesamt können Fehler unterschiedlicher Art entstehen.

Fehler 1. Art: Schrift wird nicht gefunden
Fehler 2. Art: Nichtschrift wird als Schrift interpretiert
Fehler 3. Art: Schrift wird gefunden, aber falsch gelesen

Das oben angegebene Verfahren kann insbesondere Fehler der 1. und 2. Art verringern. Wenn die Vermutung besteht, dass es sich um Druckschrift handelt (blaue Färbung in Abb. 17), kann die ausgeschnittene Textzeile an eine OCR-Engine weitergereicht werden. Diese bestätigt durch die Erkennungsprozedur die Vermutung oder verwirft sie. Handschriftliche Zeilen können an eine Handschriften- oder Schreibererkennung weitergeleitet werden, um z.B. den Autor der Schrift zu bestimmen. An Verfahren, die dieses leisten, wird z.Z. mit schon sehr guten Resultaten gearbeitet.

4 Literaturverzeichnis

SOBOTTKA, K., BUNKE, H., KRONENBERG, H., Identification of text on colored book and journal covers, Proceedings of the 5. Int. Conference on Document Analysis and Recognition 1997

KANUNGO, T., What fraction of images on the web contain text, Proceedings of Web Document Analysis, 2001

WU, V., MANMATHA, R., RISEMAN, E. M., Finding text in images, Proc. ACM Int. Conf. Digital Libraries 1997

LIENHART, R., STUBER, F., Automatic text recognition in digital videos, Proceedings of the SPIE Image and Video Processing IV 1996

CHEN, X., YUILLE, A., Detecting and Reading Text in Natural Scenes, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004

STEINKE, K.-H., Recognition of Writers by Handwriting Images; Conference on Pattern Recognition, 1980, Oxford, published in Pattern Recognition 1981; M. Duff Ed.

STEINKE, K.-H., DZIDO, R., GEHRKE, M., PRÄTEL, K., Feature recognition for herbarium specimens (Herbar-Digital), Proceedings of TDWG, Perth, 2008